Error Exponents for Channel Coding with Application to Signal Constellation Design

Jianyi Huang *Member, IEEE,* and Sean Meyn *Fellow, IEEE,* and Muriel Médard *Member, IEEE*

Abstract— **This paper concerns error exponents and the structure of input distributions maximizing the random coding exponent for a stochastic channel model. The following conclusions are obtained under general assumptions on the channel statistics:**

(i) The optimal distribution has a finite number of mass points, or in the case of a complex channel, the amplitude has finite support.

(ii) A new class of algorithms is introduced based on the cutting-plane method to construct an optimal input distribution. The algorithm constructs a sequence of discrete distributions, along with upper and lower bounds on the random coding exponent at each iteration.

(iii) In some numerical example considered, the resulting code significantly out-performs traditional signal constellation schemes such as QAM and PSK for all rates below the capacity.

Index Terms— **Information theory, channel coding, error exponents, fading channels.**

I. Introduction

The problem of constellation design has recently received renewed attention in information theory and communication theory. While many techniques in information theory such as coding have readily found their way into communication applications, the signal constellations that information theory envisages and those generally considered by practitioners differ significantly. In particular, while the optimum constellation for an additive Gaussian noise (AWGN) channel is a continuous constellation that allows for a Gaussian distribution on the input, commonly used constellations over AWGN channels, such as quadrature amplitude modulation (QAM), are not only discrete, but also generally regularly spaced. This gap between theory and practice can be explained in part by the difficulty of deploying, in practical systems, continuous constellations.

However, there is also a body of work that strongly suggests the continuous paradigm favored by theoreticians is inappropriate for realistic channel models in the majority of today's applications, such as wireless communication systems. Under any of the following conditions the optimal capacity achieving distribution has a finite number of mass points, or in the case of a complex channel, the amplitude has finite support:

(i) The AWGN channel under a peak power constraint [54], [51], [46], [16].

(ii) Channels with fading, such as Rayleigh [1] and Rician fading [34], [33]. Substantial generalizations are given in [37], [36].

(iii) Lack of channel coherence [40]. For the noncoherent Rayleigh fading channel, a Gaussian input is shown to generate bounded mutual information as SNR goes to infinity [17], [55].

(iv) Under general conditions a binary distribution is optimal, or approximately optimal for sufficiently low SNR ([29], and [56, Theorem 3].)

The finiteness of the support of the optimal input distribution bodes well for implementation of optimal constellations in communication systems, but the matter of the specific choice of points remains. The problem of selecting such a constellation is one of nonlinear optimization when channel capacity is considered [13], [30], [9], [37], [18]. While capacity provides a fundamental characterization of channel performance, the issue of how to achieve rates close to capacity with low probability of error can also be characterized in a rigorous fashion through the use of error exponents. In this paper, we pose the constellation design problem in the context of error exponent optimization.

We consider a stationary, memoryless channel with input alphabet X, output alphabet Y, and transition density defined by

$$\mathsf{P}(Y \in dy \mid X = x) = p(y|x)\, dy\,, \qquad x \in \mathsf{X},\ y \in \mathsf{Y}. \tag{1}$$

It is assumed that Y is equal to either $\mathbb{R}$ or $\mathbb{C}$, and we assume that X is a closed subset of $\mathbb{R}$. Let $\mathcal{M}$ denote the set of probability measures on the Borel σ-field $\mathcal{B}(\mathsf{X})$. For a given input distribution $\mu \in \mathcal{M}$, the resulting output density is denoted

$$p_\mu(y) = \int \mu(dx) p(y|x), \qquad y \in \mathsf{Y}. \tag{2}$$

Throughout the paper we restrict to *noncoherent* channels in which neither the transmitter nor the receiver knows the channel state.

Complex channel models in which X is equal to $\mathbb{C}$ are treated by viewing μ as the distribution of the amplitude of the input. In this case the channel input is denoted U and the output V, with $U \in \mathsf{U} =$ a closed subset of $\mathbb{C}$, $V \in \mathsf{V} = \mathbb{C}$. It is always assumed that U is circularly symmetric, and that the

[0] J. H. is with Marvell Technology, 5488 Marvell Lane, Santa Clara, CA jianyih@marvell.com
S. M. is with the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the University of Illinois, 1308 W. Main Street, Urbana, IL 61801, URL http://black.csl.uiuc.edu:80/∼meyn meyn@uiuc.edu.
S. M. supported in part by the National Science Foundation through ITR 00-85929 and ECS 02 17836.
M. M. is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology medard@mit.edu. M. M. is supported in part by NSF grant Career 6891730.

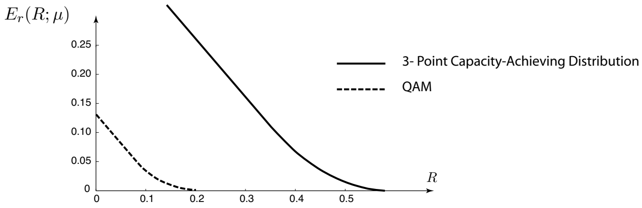

Fig. 1: The random coding exponent $E_r(R)$ for the two input distributions shown in Figure 2. The 3-point constellation performs better than 16-point QAM for all rates $R \le C$.

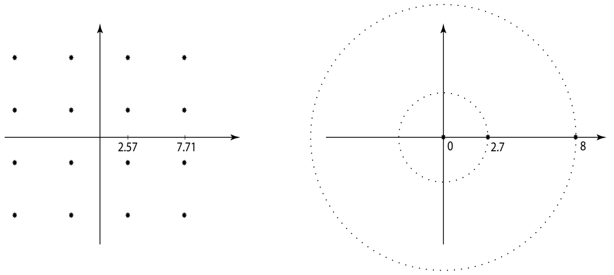

Fig. 2: The plot at left shows the 16-point QAM signal constellation, and at right is shown a three-point constellation. The respective probabilities are uniform for the QAM code, and given by $(0.5346, 0.1379, 0.397)$ for the respective codewords in the three-point constellation.

transition density on $\mathbb{C} \times \mathbb{C}$ satisfies the symmetry condition,

$$p_\bullet(v|u) = p_\bullet(e^{j\alpha}v|e^{j\alpha}u), \qquad u, v \in \mathbb{C},\ \alpha \in \mathbb{R}. \tag{3}$$

Under (3) we define,

(i) $X = |U|$ and $\mathsf{X} = \mathsf{U} \cap \mathbb{R}_+$.

(ii) For any $\mu \in \mathcal{M}$, we define $\mu_\bullet$ as the symmetric distribution on $\mathbb{C}$ whose magnitude has distribution μ. That is, we have the polar-coordinates representation: for $x > 0$, $0 \le \alpha \le 2\pi$,

$$\mu_\bullet(dx \times d\alpha) = \frac{1}{2\pi x}\mu(dx)d\alpha\,, \tag{4}$$

and $\mu(\{0\}) = \mu_\bullet(\{0\})$.

(iii) The transition density $p(\,\cdot\,|\,\cdot\,)$ on $\mathbb{C} \times \mathbb{R}_+$ is defined by

$$p(y|x) := \frac{1}{2\pi}\int_0^{2\pi} p_\bullet(y|xe^{j\theta}), \qquad y \in \mathbb{C},\ x \in \mathbb{R}_+. \tag{5}$$

Symmetry is a natural assumption in many special cases, such as Rayleigh and Rician channels. When this condition holds, it follows from [37, Proposition 2.4] that the input distribution maximizing mutual information can be assumed symmetric without loss of generality. A similar result is obtained for the error exponent in Proposition 1.1 below.

For any rate $R > 0$ the *channel reliability function* is,

$$E(R) = \lim_{N\to\infty}\Big[-\frac{1}{N}\log p_e(N,R)\Big], \qquad R > 0,$$

where $p_e(N,R)$ is the minimal probability of error, over all *block codes* of length N and rate R. If one can design a distribution with a large error exponent, then the associated random code can be constructed with a correspondingly small block-length. This has tremendous benefit in implementation.

The main goal in this paper is to maximize $E(R)$ for a given rate R, subject to two linear constraints:

(i) The *average power constraint* that

$$\langle \mu, \phi\rangle \le \sigma_P^2$$

where $\langle \mu, \phi\rangle := \int \phi(x)\ \mu(dx)$, and $\phi(x) := x^2$ for $x \in \mathbb{R}$.

(ii) The *peak power constraint* that μ is supported on $\mathsf{X} \cap [-M, M]$ for a given $M \le \infty$.

The constraints on the input distribution are expressed $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, where $\mathcal{M}(\sigma_P^2, M, \mathsf{X}) \subset \mathcal{M}$ is defined by,

$$\mathcal{M}(\sigma_P^2, M, \mathsf{X}) := \big\{\mu : \langle\mu,\phi\rangle \le \sigma_P^2,\ \mu\{[-M,M]\} = 1\big\}. \tag{6}$$

To construct a tractable optimization problem we focus on the *random coding exponent* $E_r(R)$ rather than the channel reliability function, where for a given $R \ge 0$,

$$E_r(R) := \sup_{0\le\rho\le1}\big[-\rho R + \sup_{\mu\in\mathcal{M}(\sigma_P^2,M,\mathsf{X})}(-\log(G^\rho(\mu)))\big] \tag{7}$$

$$\textit{with}\quad G^\rho(\mu) := \int\Big[\int \mu(dx)p(y|x)^{1/(1+\rho)}\Big]^{1+\rho}dy\,. \tag{8}$$

This is just one representation, based on [11, Theorem 18]. Other representations are given below in Section II-A.3. The *random-coding bound* holds under the assumptions imposed in this paper,

$$p_e(N,R) \le \exp(-NE_r(R)), \qquad N \ge 1, R \ge 0.$$

Moreover, the equality $E(R) = E_r(R)$ holds for rates greater than the *critical rate* R_{crit} [13], [30].

The random coding exponent serves as an important means of establishing the trade-off between code length and codeword error probability both for channel coding [30], [53], [4], [26], [25], [31] and source coding [38], [22]. The existence of codes providing both positive error exponents and decodability in polynomial time was established early [25]. Explicit code constructions were also available early, since low-density parity check codes themselves have exponential decay in error with code length [28], [27]. However, minimum distance rather than error exponent was, for a long time, the predominant measure of performance. Recently, the virtual rediscovery of LPDCs and the advent of further codes with a pseudo-random structure has provided a further impetus to considering error exponents for codes such as Turbo codes [39] or expander codes [7], [8]. This phenomenon is well characterized in [6]: “The performance of random codes is one of the earliest topics in information theory, dating back to Shannon’s random code ensemble (RCE). Our interest in this topic has been reawakened recently by the development of ‘random-like’ capacity-approaching codes.”

To illustrate the *dramatic* improvements that can be obtained using a carefully designed signal constellation we consider the normalized Rayleigh channel. This is the complex channel model $V = AU + N$, with A, N, each complex Gaussian, circularly symmetric, and mutually independent with $\sigma_A^2 = 1$ and $\sigma_N^2 = 1$. It is shown in [1] that the transition density for the corresponding real channel is given by,

$$p(y|x) = \frac{1}{1+x^2}\exp\Big(-\frac{1}{1+x^2}\,y\Big), \qquad x, y \in \mathbb{R}_+. \tag{9}$$

Shown at left in Figure 2 is the signal constellation for 16-point QAM. The input distribution on this constellation is uniform, which corresponds to the average power $\sigma_P^2 = 26.4$. Shown at right in Figure 2 is the input distribution constructed in [37] that achieves the maximal capacity over all distributions with average power constraint $\sigma_P^2 = 26.4$ and peak power constraint $|X| \le 8$. The mutual information is about $C = 0.5956$ nats/symbol for the optimal input distribution, which is about 3 times larger than the mutual information achieved by 16-point QAM.

With these two distributions fixed, we computed the error exponent,

$$E_r(R;\mu) := \sup_{0\le\rho\le 1} [-\rho R - \log(G^\rho(\mu))]$$

for values of $R < C$. Although the three point distribution does not maximize the error exponent for each R, the results in Figure 1 show significant performance improvement when compared with QAM for all rates less than the channel capacity. These results are consistent with other numerical results presented here and in [37], [36]. Typically, an optimal code has simple structure, yet it can *significantly* out-perform traditional signal constellation schemes such as QAM or PSK.

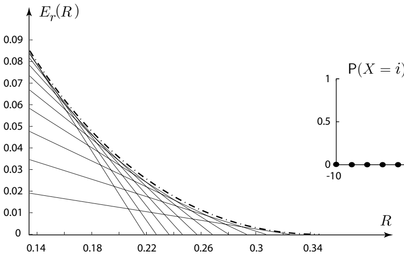

Fig. 3: Random coding exponent $E_r(R)$ for the real AWGN channel, expressed as the maximum of supporting linear functions. These numerical results were obtained using the cutting plane algorithm introduced in Section III.

Optimization of the random coding exponent is addressed as follows. Rather than parameterize the optimization problem by the given rate $R > 0$, we consider for each $\rho \in (0, 1]$ the convex program,

$$\begin{array}{rl} \textbf{inf} & G^\rho(\mu) \\ \textbf{subject to} & \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}). \end{array} \tag{10}$$

Corresponding to the minimum value $G^{\rho,*}$ is the line in the "E_r–R plane" given by,

$$L_\rho(R) := -\rho R - \log(G^{\rho,*}), \qquad R > 0.$$

Equation (7) asserts that $E_r(R)$ is equal to the maximum of these lines over $\rho \in [0, 1]$, evaluated at the given rate R. Figure 3 illustrates this representation in a numerical experiment using the AWGN channel. The parameters used in this experiment are given in Section III.

For a complex channel the expressions for $E_r(R)$ and $G^\rho(\mu)$ are identical to those given in (7, 8). However, for a symmetric channel satisfying (3) we instead consider the convex program (10) based on the real channel (5), which is justified by Proposition 1.1. The proof follows directly from convexity of G^ρ, exactly as in the proof of the corresponding result [37, Proposition 2.4].

Proposition 1.1: Suppose that (3) holds for a complex channel, and that (A1)–(A3) hold for the real channel with transition density given in (5). Given any input distribution $\mu_\bullet$ for the complex channel with transition density $p_\bullet$, let $\mu_\bullet^\circ$ denote the symmetric distribution defined by,

$$\mu_\bullet^\circ\{A\} := \frac{1}{2\pi}\int_0^{2\pi} \mu_\bullet^\circ\{Ae^{j\theta}\}\, d\theta, \qquad A \in \mathcal{B}(\mathbb{C}).$$

Then $G^\rho(\mu_\bullet^\circ) \le G^\rho(\mu_\bullet)$ for any ρ. □

Instead of studying individual channel models, which have been the topics of numerous papers, we take a systematic approach to study these problems under very general assumptions on the channel. Many common channel models such as Rician, Rayleigh and phase-noise channel models fall into these settings as shown by examples in [37]. We believe this viewpoint will clarify our applications of optimization theory to information theory.

The theory and algorithms developed in this paper are based on the *error exponent sensitivity function*, defined for $x \in \mathbb{R}$ by

$$g_\mu^\rho(x) := \int \Big[\int \mu(dz) p(y|z)^{1/(1+\rho)}\Big]^\rho p(y|x)^{1/(1+\rho)}\, dy. \tag{11}$$

The objective function in (10) can be written $G^\rho(\mu) = \langle \mu, g_\mu^\rho\rangle$. Similarly, mutual information can be expressed $I(\mu) = \langle \mu, g_\mu\rangle$ where the *channel sensitivity function* is defined by

$$g_\mu(x) := \int p(y|x)\log\big[p(y|x)/p_\mu(y)\big]\, dy\,, \qquad x \in \mathbb{R}, \tag{12}$$

where p_μ is defined in (2) as the marginal of Y. The following limit follows from L'Hôpital's rule and elementary calculus,

$$-g_\mu(x) = \lim_{\rho\to 0}\frac{\log g_\mu^\rho(x)}{\rho}, \qquad x \in \mathbb{R}.$$

The sensitivity function g_μ is easily computed numerically for the Rayleigh or phase-noise channels (see [36].) For the general Rician model, computation of g_μ appears to be less straightforward since this requires computation of $g_{\mu_\bullet}$, which involves integration over the complex plane.

The existence of a solution to (10) requires some conditions on the channel and its constraints. We list here the remaining assumptions imposed on the real channel in this paper.

(A1) The input alphabet X is a closed subset of $\mathbb{R}$, $\mathsf{Y} = \mathbb{C}$ or $\mathbb{R}$, and $\min(\sigma_P^2, M) < \infty$.

(A2) Y is large when X is large: For each $n \ge 1$,

$$\lim_{|x|\to\infty} \mathsf{P}(|Y| < n \mid X = x) = 0$$

(A3) The function $\log(p(\,\cdot\,|\,\cdot\,))$ is continuous on $\mathsf{X}\times\mathsf{Y}$ and, for any $y \in \mathsf{Y}$, $\log(p(y|\,\cdot\,))$ is analytic within the interior of X. Moreover, g_μ and g_μ^ρ are analytic functions on $\mathbb{R}$ for any $\rho \in (0, 1]$ and $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$.

We occasionally also assume,

(A4) For each ρ, if $\mu^0 \ne \mu^1$ then, for *some* y, $\int[\mu^0(dz) - \mu^1(dz)]p(y|z)^{1/(1+\rho)} \ne 0$.

Conditions (A1)-(A3) are also the standing assumptions in [37]. It is shown there that these assumptions are satisfied in all of the standard models, including the AWGN, phase-noise, Rayleigh, and Rician channels.

It is shown in Proposition 2.5 that the functional G^ρ is strictly convex under (A4).

The capacity-achieving input distribution is discrete under the conditions imposed here when M is finite. The following result is taken from [37].

Theorem 1.2: Consider a complex channel model satisfying the symmetry condition (3), whose transition density defined in (5) satisfies Assumptions (A1)–(A3), with $M < \infty$. Then, there exists an optimizer $\mu^*_\bullet$ of the convex program defining capacity,

$$\begin{array}{rl} \textbf{sup} & I(\mu) \\ \textbf{subject to} & \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}), \end{array} \tag{13}$$

with $\mu^\rho_\bullet$ symmetric, and with magnitude μ^ρ possessing finite support. □

We prove in this paper that the distribution optimizing the error exponent $E_r(R)$ is *always* discrete in the real channel, with or without a peak power constraint. In the symmetric complex channel, the distribution is symmetric and the distribution of its magnitude has finite support. The following result provides a summary of results for a symmetric complex channel. The proof follows from Propositions 1.1, 2.7 and 2.8.

Theorem 1.3: Consider a complex channel model satisfying the symmetry condition (3), whose transition density p defined in (5) satisfies Assumptions (A1)–(A3). Then, for the channel model with density p,

(i) For each ρ there exists an optimizer μ^ρ achieving the minimal value $G^{\rho,*}$. The optimizer has finite support in $\mathbb{R}_+$.

(ii) For each $R \in (0, C)$, where C denotes the value of (13), there exists ρ^* achieving the supremum in (7) so that,

$$E_r(R) = -\rho^* R - \log(G^{\rho^*,*}) = -\rho^* R - \log(G^{\rho^*}(\mu^{\rho^*}))\,.$$

(iii) The value $E_r(R)$ obtained in (ii) is also the random coding exponent for the complex channel, and the symmetric distribution with magnitude μ^{ρ^*} achieves $E_r(R)$ for the complex channel.

□

The Blahut-Arimoto (B-A) algorithm is a well-known method for computing channel capacity [12], and a related algorithm can be used to compute the random coding exponent [3]. Evidence is presented in [44] that the B-A algorithm has linear convergence, and based on a comparison with the gradient algorithm they obtain a new algorithm which also has linear convergence, but the convergence is faster than B-A in the examples considered. Recently, in [37] we have introduced a specialized algorithm based on the fact that the optimizing input distribution is discrete. One version of the algorithm can be implemented on channels with continuous input and output alphabets. The convergence is remarkably fast in all of the examples considered.

The *cutting plane algorithms* introduced in Section III are an extension of these algorithms to compute E_r along with the optimal input distribution. The fast convergence observed in [37] is also seen in all of the examples considered in this paper and in [36] for computation of E_r.

The remainder of the paper is organized as follows: Section II reviews recent as well as classical results from hypothesis testing, along with results from [37] showing that the capacity-achieving input distribution is discrete. Analogous results are then established for the distribution optimizing the error exponent. Based on this structure, algorithms are proposed in Section III to compute the optimal input distribution. Section IV concludes the paper.

II. Convex Optimization and Channel Coding

The focus of this section is to characterize optimal input distributions in three central areas of information theory: hypothesis testing, channel capacity, and error exponents. One foundation of this section lies in the theory of convex optimization [10], [15]. In particular, the structural properties obtained for optimal input distributions are based on convex duality theory and the Kuhn-Tucker alignment conditions. Related approaches to duality theory for error exponents is contained in [9], [18], [42].

A second foundation of this section is entropy. Recall that for two distributions $\mu, \pi \in \mathcal{M}$, the relative entropy, or Kullback-Leibler divergence is defined as,

$$D(\mu \parallel \pi) = \begin{cases} \langle \mu, \log \frac{d\mu}{d\pi} \rangle & \text{if } \mu \prec \pi, \\ \infty & \text{otherwise,} \end{cases}$$

where $\langle \mu, \log \frac{d\mu}{d\pi} \rangle := \int \mu(dx) \log \frac{d\mu}{d\pi}$ and the notation $\mu \prec \pi$ means that μ is absolutely continuous with respect to π, so that $\mu(dx) = r(x)\pi(dx)$, with r equal to the Radon-Nikodym derivative $r = d\mu/d\pi$. Relative entropy plays a fundamental role in hypothesis testing and communications, and it arises as the natural answer to several important questions in applications in data compression, model-selection in statistics, and signal processing [41], [23], [19], [9], [5], [20], [21], [32], [50], [14], [24], [43].

A. Hypothesis testing and reliable communication

In Section II-A.1 we survey some results from [35], [11], [57] on asymptotic hypothesis testing based on Sanov's Theorem. These results will be used to set the stage for the convex analytic methods and geometric intuition to be applied in the remainder of the paper.

We first recall Sanov's Theorem: If $\boldsymbol{X}$ is a real-valued sequence, the empirical distributions $\{\Gamma_N : N \geq 1\}$ are defined as the sequence of discrete probability distributions on $\mathcal{B}(\mathsf{X})$,

$$\Gamma_N(A) = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}\{X_k \in A\}, \qquad A \in \mathcal{B}(\mathsf{X}). \tag{14}$$

Suppose that $\boldsymbol{X}$ is i.i.d. with marginal distribution π. Sanov's Theorem states that for any closed convex set $\mathcal{A} \subseteq \mathcal{M}$,

$$\lim_{N\to\infty} -N^{-1} \log \mathsf{P}\{\Gamma_N \in \mathcal{A}\} = \inf\{D(\mu\|\pi) : \mu \in \mathcal{A}\}.$$

The relative entropy is jointly convex on $\mathcal{M} \times \mathcal{M}$, and hence computation of the minimum of $D(\mu\|\pi)$ amounts to solving a convex program.

1) Neyman-Pearson Hypothesis Testing: Consider the binary hypothesis testing problem based on a finite number of observations from a sequence $\boldsymbol{X} = \{X_t : t = 1, \ldots\}$, taking values in the set $\mathsf{X} = \mathbb{R}^d$. It is assumed that, conditioned on the hypotheses H_0 or H_1, these observations are independent and identically distributed (i.i.d.). The marginal probability distribution on X is denoted π^j under hypothesis H_j for $j = 0, 1$. The goal is to classify a given set of observations into one of the two hypotheses.

For a given $N \geq 1$, suppose that a decision test ϕ_N is constructed based on the finite set of measurements $\{X_1, \ldots, X_N\}$. This may be expressed as the characteristic function of a subset $A_1^N \subset \mathsf{X}^N$. The test declares that hypothesis H_1 is true if $\phi_N = 1$, or equivalently, $(X_1, X_2, \ldots, X_N) \in A_1^N$. The performance of a *sequence* of tests $\boldsymbol{\phi} := \{\phi_N : N \geq 1\}$ is reflected in the error exponents for the type-I error probability and type-II error probability, defined respectively by,

$$\begin{aligned} J_\phi &:= -\liminf_{N\to\infty} \frac{1}{N} \log(\mathsf{P}_{\pi^0}(\phi_N(X_1, \ldots, X_N) = 1)), \\ I_\phi &:= -\liminf_{N\to\infty} \frac{1}{N} \log(\mathsf{P}_{\pi^1}(\phi_N(X_1, \ldots, X_N) = 0)), \end{aligned} \tag{15}$$

The asymptotic N-P criterion of Hoeffding [35] is described as follows: For a given constant $\eta \geq 0$, an optimal test is the solution to the following optimization problem,

$$\beta^* = \sup\{I_\phi : \textit{ subject to } J_\phi \geq \eta\}\,, \tag{16}$$

where the supremum is over all test sequences $\boldsymbol{\phi}$.

The optimal value of the exponent I_ϕ in the asymptotic N-P problem is described in terms of relative entropy. It is shown in [57] that one may restrict to tests of the following form without loss of generality: for a closed set $\mathcal{A} \subseteq \mathcal{M}$ containing π^1,

$$\phi_N = \mathbb{I}\{\Gamma_N \in \mathcal{A}\}, \tag{17}$$

where $\{\Gamma_N\}$ denotes the sequence of empirical distributions defined in (14). Sanov's Theorem tells us that for any test of this form,

$$\begin{aligned} I_\phi &= \inf\{D(\gamma\|\pi^1) : \gamma \in \text{closure}\,(\mathcal{A}^c)\}, \\ J_\phi &= \inf\{D(\gamma\|\pi^0) : \gamma \in \mathcal{A}\}. \end{aligned} \tag{18}$$

For an arbitrary measure $\pi \in \mathcal{M}$ and for $\beta \in \mathbb{R}_+$, consider the *divergence set*,

$$\mathcal{Q}_\beta^+(\pi) := \{\gamma \in \mathcal{M} : D(\gamma \parallel \pi) \leq \beta\}. \tag{19}$$

The divergence set $\mathcal{Q}_\beta^+(\pi)$ is a closed convex subset of $\mathcal{M}$ since $D(\cdot \parallel \cdot)$ is jointly convex and lower semi-continuous on $\mathcal{M} \times \mathcal{M}$. Consequently, applying (18), the smallest closed set $\mathcal{A}$ that gives $J_\phi \geq \eta$ is the divergence set $\mathcal{A}^* = \mathcal{Q}_\eta^+(\pi^0)$, and the solution β^* to (16) is the value of the convex program, $\beta^* =$

$$\sup\{\beta : \mathcal{Q}_\eta^+(\pi^0) \cap \mathcal{Q}_\beta^+(\pi^1) = \emptyset\} = \inf_{\gamma \in \mathcal{Q}_\eta^+(\pi^0)} D(\gamma \parallel \pi^1). \tag{20}$$

Theorem 2.1 may be interpreted geometrically as follows. We have $\gamma^* \in \mathcal{Q}_\eta^+(\pi^0) \cap \mathcal{Q}_{\beta^*}^+(\pi^1)$, and the convex sets $\mathcal{Q}_\eta^+(\pi^0)$ and $\mathcal{Q}_{\beta^*}^+(\pi^1)$ are *separated* by the following set, which corresponds to the test sequence (17) obtained using the region (24):

$$\mathcal{H} = \{\gamma \in \mathcal{M} : \langle \gamma, \log \ell\rangle = \langle \gamma^*, \log \ell\rangle\}, \tag{21}$$

where ℓ denotes the Radon-Nikodym derivative (or likelihood ratio),

$$\ell(x) = \frac{d\pi^0(dx)}{d\pi^1(dx)}, \qquad x \in \mathbb{R}^d. \tag{22}$$

This geometry is illustrated in Figure 4.

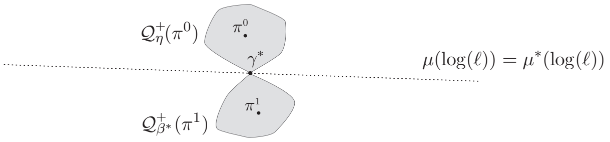
$\mathcal{Q}_\eta^+(\pi^0)$
π^0
γ^*
$\mu(\log(\ell)) = \mu^*(\log(\ell))$
π^1
$\mathcal{Q}_{\beta^*}^+(\pi^1)$

Fig. 4: The Neyman-Pearson hypothesis testing problem. The likelihood ratio test (17) with $\mathcal{A}$ defined in (24) is interpreted as a separating set between the convex sets $\mathcal{Q}_\eta^+(\pi^0)$ and $\mathcal{Q}_{\beta^*}^+(\pi^1)$.

Theorem 2.1: Suppose that $\{\pi^0, \pi^1\}$ have strictly positive densities on $\mathsf{X} = \mathbb{R}^d$, denoted $\{p^0, p^1\}$, and suppose that the optimal value of I_ϕ in (16) is finite and non-zero. Then the following statements hold,

(i) The optimal value of (16) is given by the minimal Kullback-Leibler divergence β^* given in (20).

(ii) There exists $\rho^* > 0$ such that the following alignment condition holds for the optimizer γ^* achieving the infimum in (20):

$$\log \frac{d\gamma^*}{d\pi^1}(x) + \rho^* \log \frac{d\gamma^*}{d\pi^0}(x) \leq \beta^* + \rho^*\eta, \qquad x \in \mathsf{X},$$

with equality almost everywhere. Consequently, the optimizer $\gamma^* \in \mathcal{Q}_\eta^+(\pi^0)$ has density,

$$q^*(x) = k_0[p^0(x)]^{\frac{\rho^*}{1+\rho^*}}[p^1(x)]^{\frac{1}{1+\rho^*}}, \qquad x \in \mathsf{X}, \tag{23}$$

where $k_0 > 0$ is a normalizing constant.

(iii) We have $\beta^* =$

$$\max_{\rho \geq 0}\Big\{-\rho\eta - (1+\rho)\log\Big(\int (p^0(x))^{\frac{\rho}{1+\rho}}(p^1(x))^{\frac{1}{1+\rho}}\,dx\Big)\Big\},$$

where the maximum is attained at the value of ρ^* in (ii).

(iv) The log-likelihood ratio test (LRT) is optimal, described as the general test (17) using the set,

$$\mathcal{A} := \big\{\gamma \in \mathcal{M} : \langle \gamma, \log \ell\rangle \ \leq\ \beta^* - \eta\big\}, \tag{24}$$

where ℓ denotes the likelihood ratio (22).

Proof: Part (i) of is due to Hoeffding [35], and Parts (ii) and (iii) were established in [11].

Part (iv) follows from the geometry illustrated in Figure 4: As described above, the set $\mathcal{H}$ defined in (21) defines a separating set between the convex sets $\mathcal{Q}_\eta^+(\pi^0)$ and the set $\mathcal{A}$ containing $\mathcal{Q}_{\beta^*}^+(\pi^1)$ since

$$\langle \gamma^*, \log \ell\rangle = \langle \gamma^*, \log \frac{d\gamma^*}{d\pi^1}\rangle - \langle \gamma^*, \log \frac{d\gamma^*}{d\pi^0}\rangle = \beta^* - \eta.$$

□

In typical applications it is unrealistic to assume precise values are known for the two marginals π^0, π^1. Consider the following relaxation in which hypothesis H_i corresponds to the assumption that the marginal distribution lies is a closed, affine subset $\mathbb{P}_i \subset \mathcal{M}$. A robust N-P hypothesis testing problem is formulated in which the worst-case type-II exponent is maximized over $\pi^1 \in \mathbb{P}_1$, subject to a uniform constraint on the type-I exponent over all $\pi^0 \in \mathbb{P}_0$:

$$\sup_{\phi} \inf_{\pi^1 \in \mathbb{P}_1} I_\phi^{\pi^1} \qquad \text{subject to} \qquad \inf_{\pi^0 \in \mathbb{P}_0} J_\phi^{\pi^0} \geq \eta. \tag{25}$$

A test is called optimal if it solves this optimization problem.

The optimization problem (25) is considered in [48], [49] in the special case in which the uncertainty sets are defined by specifying a finite number of generalized *moments*: A finite set of real-valued continuous functions $\{f_j : j = 1, \dots, n\}$ and real constants $\{c_j^i : j = 1, \dots, n\}$ are given, and

$$\mathbb{P}_i := \{\pi \in \mathcal{M} : \langle \pi, f_j \rangle = c_j^i, \quad j = 0, \dots, n\}, \qquad i = 0, 1. \tag{26}$$

As a notational convenience we take $f_0 \equiv 1$ and $c_0^1 = 1$.

It is possible to construct a simple optimal test based on a linear function of the data. Although the test itself is not a log-likelihood test, it has a geometric interpretation that is entirely analogous to that given in Theorem 2.1. Define for any closed set $\mathbb{P} \subset \mathcal{M}$ and $\beta > 0$ the divergence set $\mathcal{Q}_\beta^+(\mathbb{P})$:=

$$\bigcup_{\pi \in \mathbb{P}} \mathcal{Q}_\beta^+(\pi) = \bigl\{\gamma \in \mathcal{M} : D(\gamma \parallel \pi) \leq \beta \textit{ for some } \pi \in \mathbb{P}\bigr\}.$$

The value β^* in an optimal test can be expressed,

$$\beta^* = \inf\{\beta : \mathcal{Q}_\eta^+(\mathbb{P}_0) \cap \mathcal{Q}_\beta^+(\mathbb{P}_1) \neq \emptyset\}. \tag{27}$$

Moreover, the infimum is achieved by some $\mu^* \in \mathcal{Q}_\eta^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$, along with *least favorable* distributions $\pi^{0*} \in \mathbb{P}_0$, $\pi^{1*} \in \mathbb{P}_1$, satisfying

$$D(\mu^* \| \pi^{0*}) = \eta, \quad D(\mu^* \| \pi^{1*}) = \beta^*.$$

The distribution μ^* can be expressed,

$$\mu^*(dx) = \ell_0(x)\pi^{0*}(dx) = \ell_1(x)\pi^{1*}(dx)$$

where each of the functions ℓ_0, ℓ_1 is a linear combination of the constraint functions $\{f_i\}$. Each of these functions defines a separating hyperplane between the convex sets $\mathcal{Q}_\eta^+(\mathbb{P}_0)$ and $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$, as illustrated in Figure 5. Proposition 2.2 is taken from [48, Proposition 2.4]. Further results may be found in [49], [47].

Note that the function $\log \ell_0$ used in the optimal test is defined everywhere, yet in applications the likelihood rato $d\mu^*/d\pi^{0*}$ may be defined only on a small subset of X.

Proposition 2.2: Suppose that the moment classes $\mathbb{P}_0$ and $\mathbb{P}_1$ each satisfy the non-degeneracy condition that the vector $(c_0^i, \dots, c_n^i)$ lies in the relative interior of the set of all possible moments $\{\mu(f_0, \dots, f_n) : \mu \in \mathcal{M}\}$. Then, there exists $\{\lambda_0, \dots, \lambda_n\} \in \mathbb{R}$ such that the function $\ell_0 = \sum \lambda_i f_i$ is non-negative valued, and the test (17) is optimal based on the set,

$$\mathcal{A} := \bigl\{\gamma \in \mathcal{M} : \langle \gamma, \log \ell_0 \rangle \ \geq \ \eta\bigr\}, \tag{28}$$

□

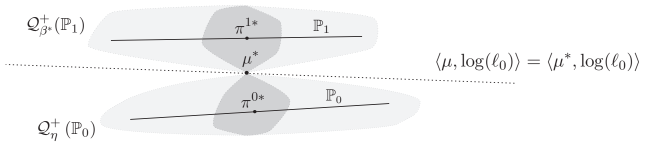
$\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$
π^{1*}
$\mathbb{P}_1$
μ^*
$\langle \mu, \log(\ell_0) \rangle = \langle \mu^*, \log(\ell_0) \rangle$
π^{0*}
$\mathbb{P}_0$
$\mathcal{Q}_\eta^+(\mathbb{P}_0)$

Fig. 5: The two-moment worst-case hypothesis testing problem. The uncertainty classes $\mathbb{P}_i$, $i = 0, 1$ are determined by a finite number of linear constraints, and the thickened regions $\mathcal{Q}_\eta^+(\mathbb{P}_0)$, $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ are each convex. The linear threshold test is interpreted as a separating hyperplane between these two convex sets.

2) Mutual Information: Channel capacity can also be expressed as the solution to the nonlinear constrained convex optimization problem (13). We derive this result based on the asymptotic N-P criterion of Hoeffding, following ideas in Anantharam [2] (see also [19], [23].)

Consider the classical decoding problem in which a set of N-dimensional codewords are generated by a sequence of random variables with marginal distribution μ. The receiver is given the output sequence $\{Y_1, \dots Y_N\}$ and considers an arbitrary sequence from the code book $\{X_1^i, \dots X_N^i\}$, where i is the index in a finite set $\{1, \dots, e^{NR}\}$, where R is the rate of the code. Since $\boldsymbol{X}^i$ has marginal distribution μ, $\boldsymbol{Y}$ has marginal density p_μ defined in (2).

For each fixed i, this decision process can be interpreted as a binary hypothesis testing problem in which Hypothesis H_1 is the hypothesis that i is the true codeword, and H_0 the alternative. Equivalently, we define

H_0: $\{(X_1^i, Y_1), \dots (X_N^i, Y_N)\}$ has marginal distribution

$$\pi^0[dx, dy] := \mu \otimes p_\mu[dx, dy] := \mu(dx)p_\mu(dy).$$

H_1: $\{(X_1^i, Y_1), \dots (X_N^i, Y_N)\}$ has marginal distribution

$$\pi^1[dx, dy] := \mu \odot p\,[dx, dy] := \mu(dx)p(y|x)dy.$$

Suppose that the error exponent $\eta > 0$ is given, and an optimal Neyman-Pearson LRT test is applied. Then $J_\phi = \eta$ means that,

$$\begin{aligned} \eta &= -\lim_{N\to\infty} \frac{1}{N} \log(\mathsf{P}_{\pi^0}(\phi_N(X_1, \dots, X_N) = 1)) \\ &= -\lim_{N\to\infty} \frac{1}{N} \log\bigl(\mathsf{P}\{\text{Code word } i \text{ is accepted} \mid i \neq i^*\}\bigr), \end{aligned} \tag{29}$$

where the index i^* denotes the codeword sent.

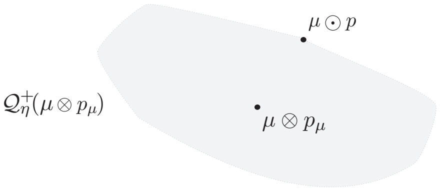

Fig. 6: The channel capacity is equal to the maximal relative entropy between $p_\mu \otimes \mu$ and $p_\mu \odot p$, over all input distributions μ satisfying the given constraints.

Consideration of e^{RN} codewords, our interest lies in the probability that at least one of the $e^{RN} - 1$ incorrect codewords is mistaken for the true codeword. We obtain through the union bound,

$$\mathsf{P}\{\text{The true codeword } i^* \text{ is rejected}\} \le \sum_{i\neq i^*} \lim_{N\to\infty} \mathsf{P}\{\text{Code word } i \text{ is accepted} \mid i \neq i^*\},$$

from which we obtain,

$$\lim_{N\to\infty} \frac{1}{N} \log\bigl(\mathsf{P}\{\text{The true code word } i^* \text{ is rejected}\}\bigr) \le R - \eta. \tag{30}$$

We must have $R < \eta$ to ensure that right hand side is negative, so that the probability that the true codeword i^* is rejected vanishes as $N \to \infty$.

One must also ensure that η is not too large, since it is necessary that the type II error exponent β^* is strictly positive so that $I_\phi > 0$ under the LRT. Hence an upper bound on R is the supremum over η satisfying $\beta^* > 0$, which is precisely mutual information:

$$I(\mu) := D(\mu \odot p \| \mu \otimes p_\mu) = \iint \log\Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)\, \mu(dx) p(y|x) dy. \tag{31}$$

This conclusion is illustrated in Figure 6.

The channel capacity is defined to be the maximum of I over all input distributions μ satisfying the given constraints. We thus arrive at the convex program (13).

3) Error exponents: A representation of the channel-coding random coding exponent is obtained based on similar reasoning. Here we illustrate the form of the solution, and show that it may be cast as a *robust* hypothesis testing problem of the form considered in Section II-A.1.

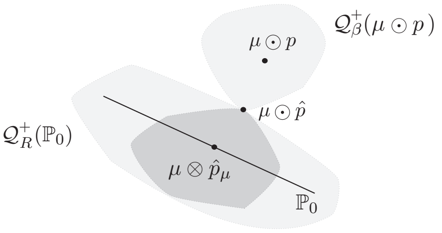

Fig. 7: The error exponent is equal to the solution of a robust N-P hypothesis testing problem.

For a given $\mu \in \mathcal{M}$, denote by $\mathbb{P}_0$ the space of product measures on $\mathsf{X} \times \mathsf{Y}$,

$$\mathbb{P}_0 = \{\mu \otimes \nu : \nu \text{ is a probability measure on } \mathsf{Y}\}, \tag{32}$$

and define the corresponding divergence set for a given $R > 0$,

$$\mathcal{Q}^+_R(\mathbb{P}_0) := \bigcup_\nu \mathcal{Q}^+_R(\mu \otimes \nu).$$

Equivalently, $\mathcal{Q}^+_R(\mathbb{P}_0) = \{\gamma : \min_\nu D(\gamma \parallel \mu \otimes \nu) \le R\}$. The robust hypothesis testing problem is binary, with H_1 as defined in the channel capacity problem, but with H_0 defined using $\mathbb{P}_0$:

H_0: $\{(X^i_j, Y_j) : j = 1, \dots, N\}$ has marginal distribution $\pi^0 \in \mathbb{P}_0$.

H_1: $\{(X^i_j, Y_j) : j = 1, \dots, N\}$ has marginal distribution $\pi^1 := \mu \odot p$.

Proposition 2.3 shows that the random coding exponent $E_r(R)$ can be represented as the solution to the robust N-P hypothesis testing problem (25) with $\eta = R$, $\mathbb{P}_0$ defined in (32), and $\mathbb{P}_1 = \{\mu \odot p\}$.

Proposition 2.3: Suppose that (A1)–(A3) hold. Then, for each rate R less than capacity the error exponent can be expressed,

$$E_r(R) = \sup_\mu \Bigl(\inf_\beta \bigl\{ \beta : \mathcal{Q}^+_\beta(\mu \odot p) \cap \mathcal{Q}^+_R(\mathbb{P}_0) \neq \emptyset \bigr\} \Bigr). \tag{33}$$

Suppose that there exists a triple (μ^*, ν^*, γ^*) that solve (33) in the sense that

$$D(\gamma^* \parallel \mu^* \odot p) = E_r(R), \quad D(\gamma^* \parallel \mu^* \otimes \nu^*) = R.$$

Then, there exists a channel transition density $\hat p$ such that

$$\gamma^* = \mu^* \odot \hat p, \quad \nu^* = \hat p_{\mu^*}, \tag{34}$$

and the rate can be expressed as mutual information,

$$R = I(\mu^*; \hat p) := D(\mu^* \odot \hat p \parallel \mu^* \otimes \hat p_\mu).$$

Proof: Blahut in [11] establishes several representations for the random coding exponent, beginning with the following

$$E_r(R) = \sup_\mu \Bigl(\inf_{\mu \odot \hat p \in \hat{\mathcal{Q}}^+_R} D(\mu \odot \hat p \parallel \mu \odot p) \Bigr), \tag{35}$$

where the supremum is over all μ, subject to the given constraints, and the infimum is over all transition densities $\hat p$ satisfying $\mu \odot \hat p \in \hat{\mathcal{Q}}^+_R$ where

$$\hat{\mathcal{Q}}^+_R := \{\mu \odot \hat p : D(\mu \odot \hat p \parallel \mu \otimes \hat p_\mu) \le R\}.$$

The optimization problem (33) is thus a relaxation of (35) in which the distributions $\{\nu\}$ in the definition of $\mathbb{P}_0$ in (32) are constrained to be of the form $\hat p_\mu$, and the distributions $\{\gamma\}$ are constrained to be of the form $\mu \odot \hat p$ for some transition density $\hat p$.

It remains to show that these restrictions hold for any solution to (33), so that the relaxed optimization problem (33) is equivalent to (35).

For any distribution γ on $\mathcal{B}(\mathsf{X} \times \mathsf{Y})$, the two marginals are denoted,

$$\gamma_1(dx) = \gamma(dx \times \mathsf{Y}), \quad \gamma_2(dy) = \gamma(\mathsf{X} \times dy).$$

For fixed μ on $\mathcal{B}(\mathsf{X})$, denote the infimum over β in (33) by,

$$\beta^*(\mu) := \inf\bigl\{ \beta : \mathcal{Q}^+_\beta(\mu \odot p) \cap \mathcal{Q}^+_R(\mathbb{P}_0) \neq \emptyset \bigr\}. \tag{36}$$

If (ν^*, γ^*) solve (36) in the sense that

$$D(\gamma^* \parallel \mu \odot p) = \beta^*(\mu), \quad D(\gamma^* \parallel \mu \otimes \nu^*) = R,$$

then the distribution γ^* solves the ordinary N-P hypothesis testing problem with $\pi^0 = \mu \otimes \nu^*$ and $\pi^1 = \mu \odot p$. It then follows that the first marginal γ^*_1 is equal to μ by the representation given in Theorem 2.1 (ii). The second marginal of γ^*_2 is equal to ν^* since for any ν,

$$D(\gamma^* \parallel \mu \otimes \nu) = D(\gamma^* \parallel \mu \otimes \gamma^*_2) + D(\gamma^*_2 \parallel \nu) \ge D(\gamma^* \parallel \mu \otimes \gamma^*_2),$$

with equality if and only if $\nu = \gamma^*_2$.

In summary, the optimizer γ^* can be expressed $\gamma^* = \mu \odot \hat{p}$ for some channel density $\hat{p}$ satisfying $\nu^* = \hat{p}_\mu$. That is, (34) holds, and it follows that γ^* is feasible (and therefore optimal) for (35), which establishes the desired equivalence between the optimization problems (33) and (35). □

The solution to the optimization problem (36) is illustrated in Figure 7. The channel transition density $\hat{p}$ shown in the figure solves

$$\begin{aligned}\beta^*(\mu) &= \inf\bigl\{\beta : \mathcal{Q}^+_\beta(\mu \odot p) \cap \mathcal{Q}^+_R(\mu \otimes \hat{p}_\mu) \neq \emptyset\bigr\} \\ &= D(\mu \odot \hat{p} \parallel \mu \odot p).\end{aligned}$$

The error exponent is equal to the maximal relative entropy $\beta^*(\mu)$ over all μ, and the rate can be expressed as mutual information $R = I(\mu^*; \hat{p}) := D(\mu^* \odot \hat{p} \parallel \mu^* \otimes \hat{p}_\mu)$ where μ^* is the optimizing distribution.

B. Alignment

The Kuhn-Tucker alignment conditions are the simplest route to establishing the solution to the Neyman-Pearson hypothesis testing problem given in Theorem 2.1, as well as the alignment conditions characterizing channel capacity [37]. Here we construct the alignment conditions that characterize the solution to (10) based on the error exponent sensitivity function g^ρ_μ defined in (11).

Boundedness of the sensitivity function is central to the analysis that follows.

Theorem 2.4: The following hold under (A1)-(A3):

(i) $0 < g^\rho_\mu(x) \le 1$ for each x.

(ii) $g^\rho_\mu \to 0$ as $x \to \infty$.

(iii) Suppose that there is a peak power constraint, so that $M < \infty$. For each finite $N > 0$, the mapping $\mu \to g_\mu$ is continuous from $\mathcal{M}(\sigma^2_P, M, \mathsf{X})$ to $L_\infty[-N, N]$. That is if $\mu_n \to \mu$ weakly, with $\mu_n \in \mathcal{M}(\sigma^2_P, M, \mathsf{X})$ for all n, then

$$\lim_{n\to\infty} \sup_{|x|\le N} |g^\rho_{\mu_n}(x) - g^\rho_\mu(x)| = 0. \tag{37}$$

Proof: It prove (i), it is obvious that g^ρ_μ is positive-valued, and by Jensen's inequality we have for any x,

$$\begin{aligned} g^\rho_\mu(x) &\le \int p_\mu(y)^{\rho/1+\rho} p(y|x)^{1/1+\rho}\,dy \\ &= \int p_\mu(y) \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{1/1+\rho} dy \\ &\le \Bigl(\int p_\mu(y) \frac{p(y|x)}{p_\mu(y)}\,dy\Bigr)^{1/1+\rho} = 1. \end{aligned} \tag{38}$$

From the first inequality in (38) we obtain, for each $N \ge 1$,

$$\begin{aligned} g^\rho_\mu(x) &\le \int p_\mu(y)^{\rho/1+\rho} p(y|x)^{1/1+\rho}\,dy \\ &= \int_{|y|\le N} \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{\frac{1}{1+\rho}} p_\mu(y)\,dy \\ &\quad + \int_{|y|\ge N} \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{\frac{1}{1+\rho}} p_\mu(y)\,dy. \end{aligned}$$

The following limit follows from (A2) and the Dominated Convergence Theorem,

$$\lim_{x\to\infty} \int_{|y|\le N} \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{\frac{1}{1+\rho}} p_\mu(y)\,dy = 0.$$

Let b_N denote the normalizing constant that makes $b_N p_\mu(y) \boldsymbol{1}(|y| \ge N)$ a probability density on Y. That is,

$$b_N^{-1} := \int_{|y|\ge N} p_\mu(y)\,dy. \tag{39}$$

Another application of Jensen's inequality then gives, as $N \to \infty$,

$$\begin{aligned} &\int_{|y|\ge N} \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{\frac{1}{1+\rho}} p_\mu(y) \\ &\qquad = b_N^{-1} \int_{|y|\ge N} \Bigl(\frac{p(y|x)}{p_\mu(y)}\Bigr)^{\frac{1}{1+\rho}} b_N p_\mu(y)\,dy \\ &\qquad \le b_N^{-1} \Bigl[\int_{|y|\ge N} b_N p(y|x)\,dy\Bigr]^{\frac{1}{1+\rho}} \le b_N^{-\frac{\rho}{1+\rho}} \to 0. \end{aligned}$$

This implies the desired conclusion in (ii),

$$0 \le \lim_{x\to\infty} g^\rho_\mu(x) \le \lim_{x\to\infty} \int p_\mu(y)^{\rho/1+\rho} p(y|x)^{1/1+\rho}\,dy = 0.$$

The proof of (iii) is again based on truncation. Define for $N \ge 1$, $x \in \mathsf{X}$,

$$g^{\rho,N}_\mu(x) := \int_{|y|\le N} \Bigl[\int \mu(dz) p(y|z)^{1/(1+\rho)}\Bigr]^\rho p(y|x)^{1/(1+\rho)}\,dy.$$

Under the peak power constraint, it follows from (A3) that the functions $\{g^{\rho,N}_\mu : \mu \in \mathcal{M}(\sigma^2_P, M, \mathsf{X})\}$ are differentiable, and the derivative is uniformly bounded over μ, and over $x \in [-N, N]$. Consequently, these functions are equicontinuous on $[-N, N]$.

Suppose that $\mu_n \to \mu$ weakly as $n \to \infty$. Then the uniform convergence (37) holds for the truncated functions:

$$\sup_{|x|\le N} |g^{\rho,N}_{\mu_n}(x) - g^{\rho,N}_\mu(x)| \to 0, \qquad n \to \infty.$$

This follows directly from equicontinuity.

Finally, it follows from weak convergence that $p_{\mu_n} \to p_\mu$ pointwise, and hence in the weak topology. From the previous arguments we have the uniform bound,

$$|g^{\rho,N}_{\mu_n}(x) - g^\rho_{\mu_n}(x)| \le [b_N(\mu_n)]^{-\frac{\rho}{1+\rho}}, \qquad x \in \mathsf{X},$$

where $b_N = b_N(\mu_n)$ is defined in (39). Convergence of $\{p_{\mu_n}\}$ implies that these distributions are *tight*, which means that $b_N^{-1}(\mu_n) \to 0$ as $N \to \infty$, uniformly in n. Hence (37) follows from the uniform convergence for the truncated functions $\{g^{\rho,N}_{\mu_n}\}$. □

It can be shown that the functional $G^\rho\colon \mathcal{M} \to \mathbb{R}_+$ is convex using the representation (8). The following set of results establishes convexity and finer results based on a representation for its derivative with respect to μ. For $\mu, \mu^\circ \in \mathcal{M}$ and $\theta \in [0,1]$ we denote $\mu_\theta := (1-\theta)\mu^\circ + \theta\mu$.

Proposition 2.5: The following hold under (A1)–(A3): For any given $\mu, \mu^\circ \in \mathcal{M}(\sigma^2_P, M, \mathsf{X})$ and $\rho > 0$,

(i) $G^\rho(\mu) = \langle \mu, g^\rho_\mu \rangle$.

(ii) The functional G^ρ is convex, and can be expressed as the maximum of linear functionals,

$$G^\rho(\mu^\circ) = \max_{\mu\in\mathcal{M}} \{(1+\rho)\langle \mu^\circ, g^\rho_\mu \rangle - \rho G^\rho(\mu)\}$$

(iii) Fix $\rho \ge 0$, $\mu^\circ \in \mathcal{M}$. The first order sensitivity is given by

$$\frac{d}{d\theta} G^\rho(\mu_\theta)\Big|_{\theta=0} = (1+\rho)\langle \mu - \mu^\circ, g^\rho_{\mu^\circ}\rangle.$$

(iv) If (A4) holds then G^ρ is strictly convex for each $\rho > 0$.

Proof: Part (i) follows from the definition.

To prove (iii), we differentiate the expression for G^ρ with respect to θ,

$$G^\rho(\mu_\theta) = \int \Big[\int \mu_\theta(dz) p(y|z)^{1/(1+\rho)} \Big]^{1+\rho} dy,$$

giving

$$\begin{aligned}\frac{d}{d\theta} G^\rho(\mu_\theta) = \int (1+\rho) \Big[\int \mu_\theta(dz) p(y|z)^{1/(1+\rho)} \Big]^{\rho} \\ \times \Big[\int [\mu(dx) - \mu^\circ(dx)] p(y|x)^{1/(1+\rho)} \Big] dy \\ = (1+\rho)\langle \mu - \mu^\circ, g^\rho_{\mu_\theta}\rangle.\end{aligned}$$

Evaluating at $\theta = 0$ then gives (iii).

Convexity of G^ρ over μ, together with the expression (iii) for the derivative gives,

$$G^\rho(\mu^\circ) \ge G^\rho(\mu) + (1+\rho)\langle \mu^\circ - \mu, g^\rho_\mu \rangle.$$

This combined with the expression $\langle \mu, g^\rho_\mu\rangle = G^\rho(\mu)$ gives (ii).

To see (iv), suppose that $\mu^0 \neq \mu^1$ are given, and that for all $\theta \in [0,1]$,

$$G^\rho(\mu_\theta) := G^\rho((1-\theta)\mu^0 + \theta\mu^1) = (1-\theta)G^\rho(\mu^0) + \theta G^\rho(\mu^1).$$

Elementary calculus implies that $\int [\mu^0(dz) - \mu^1(dz)] p(y|z)^{1/(1+\rho)}$ is identically zero, violating (A4). □

Continuity of G^ρ easily follows:

Proposition 2.6: Suppose that (A1)–(A3) hold. Then, given ρ, the mapping $G^\rho \colon \mathcal{M}(\sigma_P^2, M, \mathsf{X}) \mapsto \mathbb{R}_+$ is continuous in the weak topology.

Proof: If $\mu_n \to \mu$ weakly then the distributions are tight. Hence we can truncate all of these distributions as follows. Suppose that $\varepsilon > 0$ is given, and that N is chosen so that $(a_n^N)^{-1} := \mu_n\{[-N,N]\} \ge 1 - \varepsilon$ for each n. Denoting μ_n^N the distribution on $[-N,N]$ given by $\mu_n^N(A) = a_n^N \mu_n(A \cap [-N,N])$ for measurable $A \subset \mathsf{X}$, and defining μ^N analogously, we have

$$\begin{aligned}|G^\rho(\mu_n^N) - G^\rho(\mu_n)| &\le (1-\varepsilon)^{-1}\varepsilon, \\ |G^\rho(\mu^N) - G^\rho(\mu)| &\le (1-\varepsilon)^{-1}\varepsilon.\end{aligned}$$

This follows from convexity combined with the bound $g^\rho \le 1$.

Theorem 2.4 (iii) gives $G^\rho(\mu_n^N) \to G^\rho(\mu^N)$ as $n \to \infty$, and hence

$$\limsup_{n\to\infty} |G^\rho(\mu_n) - G^\rho(\mu)| \le 2(1-\varepsilon)^{-1}\varepsilon.$$

The result follows since $\varepsilon > 0$ is arbitrary. □

For fixed ρ, the optimization problem (10) is a convex program since G^ρ is convex. Continuity then leads to existence of an optimizer. The following result summarizes the structure of the optimal input distribution. It is similar to Theorem 2.8 of [37], which required the peak power constraint $M < \infty$. We stress that this condition is not required here.

Theorem 2.7: Suppose that (A1)–(A3) hold. Then,

(i) For each $\rho \ge 0$, there exists $\mu^\rho \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ that achieves $G^{\rho,*}$.

(ii) A given distribution $\mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ is optimal if and only if there exists a real number λ_1^* and a positive real number λ_2^* such that

$$\begin{aligned} & g^\rho_{\mu^\circ}(x) \ge \lambda_1^* - \lambda_2^* x^2, \qquad x \in \mathsf{X}, \\ \text{and} \quad & g^\rho_{\mu^\circ}(x) = \lambda_1^* - \lambda_2^* x^2, \qquad a.e.\ [\mu^\circ]\end{aligned}$$

If these conditions hold, then

$$G^{\rho,*} := \min_\mu G^\rho(\mu) = G^\rho(\mu^\circ) = \frac{\lambda_1^* - \lambda_2^*\sigma_P^2}{1+\rho}.$$

(iii) In the absence of an average power constraint $\lambda_2^* = 0$.

(iv) If (A4) holds, then the optimizer μ^ρ is unique.

Proof: Existence of μ^* follows because $\mathcal{M}(\sigma_P^2, M, \mathsf{X})$ is compact in the weak topology when $\min(\sigma_P^2, M) < \infty$, and G^ρ is continuous.

The proof of (ii) is based on the Lagrangian relaxation with objective function,

$$\mathcal{L}^\rho(\mu) := G^\rho(\mu) + \lambda_2 \int (x^2 - \sigma_P^2)\, \mu(dx), \qquad \mu \in \mathcal{M}.$$

Since the functional $\mathcal{L}^\rho \colon \mathcal{M} \to \mathbb{R}$ is convex, a distribution $\mu^\circ \in \mathcal{M}$ minimizes $\mathcal{L}^\rho$ over all probability distributions if and only if the first order conditions hold: Letting $\mu_\theta := (1-\theta)\mu^\circ + \theta\mu$ for a given $\mu \in \mathcal{M}$ and $\theta \in [0,1]$, the derivative of $\mathcal{L}^\rho(\mu_\theta)$ at $\theta = 0$ must be non-negative since μ° is a local minimum.

From the foregoing, this derivative can be expressed,

$$\begin{aligned}\frac{d}{d\theta}\mathcal{L}^\rho(\mu_\theta)\Big|_{\theta=0} &= \int \big[(1+\rho) g^\rho_{\mu^\circ}(x) + \lambda_2(x^2 - \sigma_P^2)\big]\, \mu(dx) \\ &\quad - \int \big[(1+\rho) g^\rho_{\mu^\circ}(x) + \lambda_2(x^2 - \sigma_P^2)\big]\, \mu^\circ(dx).\end{aligned}$$

Specializing to $\mu = \delta_x$ for $x \in \mathsf{X}$, we find that the statement that this derivative is non-negative for each x is equivalent to the alignment conditions in (ii). □

When $\sigma_P^2 = \infty$ so that the input distribution is only subject to a peak power constraint, then the Lagrange multiplier λ_2^* is zero, and the alignment condition in Theorem 2.7 (ii) becomes,

$$g^\rho_{\mu^\circ}(x) \ge \lambda_1^*, \quad x \in \mathsf{X}, \quad \text{and} \quad g^\rho_{\mu^\circ}(x) = \lambda_1^*, \quad a.e.\ [\mu^\circ] \tag{40}$$

Shown in Figure 10 are plots of g^ρ_μ for two distributions μ_0, μ_1 with $\rho = 0.5$ for the normalized Rayleigh channel. The input distribution μ_0 violates the alignment condition (40) and hence is not optimal. The alignment condition does hold for μ_1, and we conclude that this distribution does optimize (10) over all distributions on $[0, 10]$ (without average power constraint.)

Proposition 2.8: Suppose that (A1)–(A3) hold. Then, for given ρ, any optimal input distribution μ^* achieving $G^{\rho,*}$ is discrete, with a finite number of mass points in any interval.

Proof: To see that the optimizer $g^\rho_{\mu^*}$ is discrete consider the alignment conditions. There exists a quadratic function q^* satisfying $q^*(x) \le g^\rho_{\mu^*}(x)$, with equality a.e. $[\mu^*]$. Theorem 2.4 asserts that $g^\rho_{\mu^*}$ takes on strictly positive values and vanishes at infinity. If $M = \infty$ it follows that q^* is not a

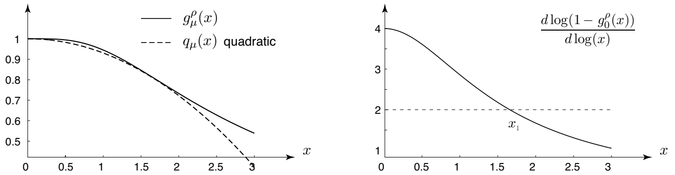

Fig. 8: Optimal binary distribution for the Rayleigh channel with $\rho = 0.5$. The plot at right shows the derivative (41) along with the point x_1 at which the derivative is equal to 2. At left is a plot of the error exponent sensitivity function g^ρ_μ for the optimal distribution supported at zero, and a point near x_1. The sensitivity function is aligned with the quadratic function shown.

constant function, and hence $q^*(x) \to -\infty$ as $x \to \infty$. This shows that the optimizer has bounded support, with

$$\mathrm{supp}\,(\mu^*) \subset \{x : q^*(x) > 0\}.$$

Moreover, since $g^\rho_{\mu^*}$ is an analytic function on X it then follows that $g^\rho_{\mu^*}(x) = q^*(x)$ is only possible for a finite number of points.

If M is finite the argument is similar: $q^*(x) = g^\rho_{\mu^*}(x)$ on the support of μ^*. If this support is infinite, then the analytic assumption implies that $q^*(x) = g^\rho_{\mu^*}(x)$ for all x, which is impossible. □

So we have shown that under our very general channel models, the optimal distribution that achieve the error exponent is always discrete. Next we show when SNR goes to zero, the optimal discrete distribution becomes binary.

C. Optimal binary distributions

We now take a closer look at the alignment conditions to establish conditions ensuring that the distribution optimizing (10) is binary for sufficiently low SNR.

Gallager in [29] bounds the random coding exponent by a linear functional over the space of probability measures. The bound is shown to be tight for low SNR, and thus the error exponent optimization problem is converted to a linear program over the space of probability measures. An optimizer is an extreme point, which is shown to be binary. Similar arguments used in [37] can be generalized to the model considered here.

We begin with consideration of *zero* SNR, which leads us to consider the senstivity function using the point mass at 0, denoted $g^\rho_0 := g^\rho_{\delta_0}(x)$. It is easy to see $g^\rho_0(0) = 1$, and we have seen that $g^\rho_0(x) \le 1$ everywhere. Given the analyticity assumption, it follows that this function has zero derivative at the origin, and non-positive second derivative. We thus obtain the bound,

$$\frac{d\log(1-g^\rho_0(x))}{d\log(x)}\Big|_{x=0} \ge 2,$$

with equality holding if and only if the second derivative of $g^\rho_0(x)$ is non-zero at $x = 0$.

Proposition 2.9 is analogous to Theorem 3.4 of [37] which establishes that the distribution achieving channel capacity is binary for sufficiently low SNR. However, unlike this previous result and the result of [29], we do not require a peak power constraint on the input distribution.

Proposition 2.9: Consider a channel with $\mathsf{X} = \mathbb{R}_+$ satisfying Assumptions (A1)–(A3). For a fixed $\rho > 0$, suppose that the following hold,

(i) $\dfrac{d^2}{dx^2} g^\rho_0\,(0) = 0.$

(ii) There is a unique $x_1 > 0$ satisfying

$$\frac{d\log(1-g^\rho_0(x))}{d\log(x)}\Big|_{x=x_1} = 2.$$

(iii) There is 'non-zero sensitivity' at x_1:

$$\frac{d}{dx}\Big(\frac{d\log(1-g^\rho_0(x))}{d\log(x)}\Big)\Big|_{x=x_1} \ne 0.$$

Then, for all SNR sufficiently small, the optimal input distribution is binary with one point at the origin.

Proof: Define $\lambda_{2,0} > 0$ via,

$$\lambda_{2,0} = \inf\{\lambda : 1 - g^\rho_0(x) \le \lambda x^2 \text{ for all } x \ge 0\},$$

and let q_0 denote the concave quadratic, $q_0(x) = 1 - \lambda_{2,0}x^2$, $x \ge 0$.

We have $g^\rho_0(x) \ge q_0(x)$ for all x, and this inequality is strict for $x > 0$ near the origin due to Assumption (i) and the fact that $g^\rho_0(x)$ has zero derivative at the origin, and non-positive second derivative. Define,

$$x^* = \inf\{x > 0 : g^\rho_0(x) = q_0(x)\}.$$

The function $\log((1-g^\rho_0(x))x^{-2})$ has a local maximum at x^*, implying its derivative is zero. Equivalently,

$$\frac{d\log(1-g^\rho_0(x))}{d\log(x)}\Big|_{x=x_1} = 2,$$

and hence $x^* = x_1$. That is, g^ρ_0 is aligned with $q_0(x) = \lambda_{0,0} - \lambda_{2,0}x^2$, with $\lambda_{0,0} = 1$.

Applying the Implicit Function Theorem (which is justified under (iii)) we can find $\varepsilon > 0$ such that for each $\sigma^2_P \in (0, \varepsilon]$, there is $x_1(\varepsilon)$ near x_1 and (λ_0, λ_2) near $(\lambda_{0,0}, \lambda_{2,0})$ such that q_{μ^*} aligned with $g^\rho_{\mu^*}$, where $q_{\mu^*} = \lambda_0 - \lambda_2 x^2$, and μ^* is the unique binary distribution supported on $\{0, x_1\}$ with second moment σ^2_P. Consequently, μ^* is optimal by Theorem 2.7. □

We now turn to the Rayleigh channel to illustrate the proof of Proposition 2.9. Given the channel transition probability function (9), the sensitivity function is,

$$g^\rho_0(x) = \frac{(1+\rho)(1+x^2)^{\frac{\rho}{1+\rho}}}{(1+x^2)\rho + 1}, \qquad x \ge 0,$$

and its log-derivative, $\frac{d\log(1-g^\rho_0(x))}{d\log(x)} =$

$$\frac{2\rho x^2(1+x^2)^{-\frac{1}{1+\rho}}[1+\rho(1+x^2)] - 2\rho x^2(1+\rho)(1+x^2)^{\frac{\rho}{1+\rho}}}{(1+\rho)[1+\rho(1+x^2)](1+x^2)^{\frac{\rho}{1+\rho}} - [1+\rho(1+x^2)]^2}. \tag{41}$$

From the plot shown at left in Figure 8 we see that there exists a quadratic function q_0 satisfying $q_0(x) \le g^\rho_0(x)$, with equality at the origin and precisely one $x_1 > 0$. The plot at right shows that (iii) holds, and hence that all of the assumptions of Proposition 2.9 are satisfied. Consequently, with vanishing SNR, the optimizing distribution is binary, and approximates the binary distribution supported on $\{0, x_1\}$.

These conclusions are very different from those obtained on considering the distribution optimizing capacity. It is known that the optimizing distribution is binary, for low SNR, with one point at the origin, but the second point of support diverges as SNR tends to zero. Similarly, one can show that g_0^ρ becomes increasingly 'flat' as $\rho \to 0$, and hence $x_1(\rho) \to \infty$ as $\rho \to 0$.

III. Algorithms

The algorithms proposed here are motivated by the discrete nature of optimal input distributions. We find in examples that the convergence is very fast.

Cutting plane algorithm The algorithm is initialized with an arbitrary distribution $\mu_0 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, and inductively constructs a sequence of distributions as follows. At the nth stage of the algorithm, we are given n distributions $\{\mu_0, \mu_1, \dots, \mu_{n-1}\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$. We then define,

(i) The piecewise linear approximation, for $\mu \in \mathcal{M}$, $\rho \geq 0$,

$$G_n^\rho(\mu) := \max_{0 \leq i \leq n-1} \{(1+\rho)\langle \mu, g_{\mu_i}^\rho \rangle - \rho \langle \mu_i, g_{\mu_i}^\rho \rangle\}. \qquad (42)$$

(ii) The next distribution,

$$\mu_n = \arg\min\{G_n^\rho(\mu) : \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})\}. \qquad (43)$$

□

The optimization problem (43) is equivalently expressed as the solution to the linear program in the variables $e \in \mathbb{R}$, $\mu \in \mathcal{M}$:

$$\begin{aligned} \textbf{min} \quad & e \\ \textbf{subject to} \quad & e \geq (1+\rho)\langle \mu, g_{\mu_i}^\rho \rangle - \rho \langle \mu_i, g_{\mu_i}^\rho \rangle, \\ & \qquad i = 0, \dots, n-1, \quad \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}). \end{aligned} \qquad (44)$$

It is evident that $G_n^\rho(\mu) \leq G^\rho(\mu)$ for all $\mu \in \mathcal{M}$. It may be shown under conditions somewhat stronger than (A1)-(A3) that the algorithm is convergent, in the sense that the sequence of distributions $\{\mu_n\}$ converges weakly to an optimal distribution. The proof is identical to the proof of [37, Theorem 4.1], based on the continuity result Proposition 2.6.

Theorem 3.1: Consider the real channel model satisfying Assumptions (A1)–(A4). The cutting plane algorithm generates a sequence of distributions $\{\mu_n : n \geq 1\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ such that

(i) $\mu_n \to \mu^*$ weakly, as $n \to \infty$;
(ii) $G^\rho(\mu_n) \to G^\rho(\mu^*)$;
(iii) $e_1 \leq e_2 \leq e_3 ... \to G^\rho(\mu^*)$;
(iv) μ_n can be chosen so that it has at most $n+1$ points of support for each $n \geq 1$.

□

Figure 3 shows results from implementation of this algorithm for a range of $\rho \geq 0$ for the real AWGN channel $Y = X + N$, with $\mathsf{X} = \{-10, -9, ..., 9, 10\}$, $\sigma_P^2 = 10$, and $\sigma_N^2 = 10$. The dashed line represents the exponent $E_r(R)$ achieved using a Gaussian input distribution. This very nearly matches the upper-envelope obtained from the solid lines obtained using the cutting plane algorithm. Hence restricting to a finite alphabet does not lead to a significant loss in performance in this example.

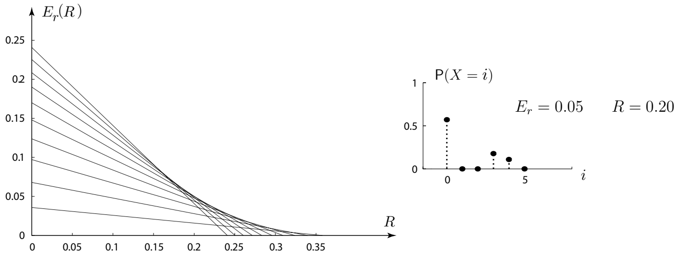

Fig. 9: Random coding exponent $E_r(R)$ obtained using the cutting plane algorithm for the Rayleigh channel.

Also shown in the figure is the input distribution achieving the maximum error exponent $E_r(R) = 0.05$, subject to μ supported on X, for the rate $R = 0.175$. The optimizer is symmetric on the four points $\{\pm 4, \pm 3\}$ with $p(X = -4) = p(X = 4) = 0.07$, $p(X = -3) = p(X = 3) = 0.43$. This is very different than the input achieving capacity which assigns positive mass to the origin in this model [37].

Results obtained for the normalized Rayleigh channel with transition density given in (9) are shown in Figure 9. In this experiment the input alphabet consisted of the six points $\mathsf{X} = \{0, 1, 2, 3, 4, 5\}$, $M = \infty$, and $\sigma_P^2 = 4$. At right is shown the distribution achieving $E_r(R) = 0.05$ at $R = 0.2$: It is supported on the three points shown, with $p(X = 0) = 0.6519$, $p(X = 3) = 0.2242$, and $p(X = 4) = 0.1239$.

Although the cutting-plane algorithm is convergent even in the infinite dimensional setting in which X is continuous, a finite-dimensional algorithm is needed in any practical application. This is the reason why the input alphabet was taken to be fixed and finite in the numerical examples illustrated here. Next, we introduce an extension of the cutting-plane method to recursively *construct* the input alphabet X.

Given a finite alphabet X_0, a sequence of finite alphabets $\{\mathsf{X}_n : n \geq 0\}$ is obtained by induction. At the nth state of the algorithm, the optimal input distribution μ_n on X_n is obtained using the cutting-plane algorithm. The detail of this procedure are described as follows.

Steepest Decent Cutting-plane algorithm The algorithm is initialized with a finite alphabet $\mathsf{X}_0 \subseteq \mathsf{X}$, together with a distribution $\mu_0 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}_0)$. At the nth stage of the algorithm, we are given n distributions $\{\mu_0, \mu_1, ..., \mu_{n-1}\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ and an input alphabet X_n. The next distribution and input alphabet are then defined as follows,

(i) The new distribution,

$$\mu_n = \arg\min\{G^\rho(\mu) : \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}_n)\}, \qquad (45)$$

where the minimization can be obtained by the cutting plane algorithm or any other convex optimization method.

(ii) The new alphabet $\mathsf{X}_{n+1} = \mathsf{X}_n \cup \{x_{n+1}\}$, where

$$x_{n+1} = \arg\min\{g_n^\rho(x) - r_n x^2 : |x| \leq M, x \in \mathsf{X}\}, \qquad (46)$$

where $g_n^\rho(x) := g_{\mu_n}^\rho$ and r_n is the associated Lagrange multiplier obtained in the solution of (45). □

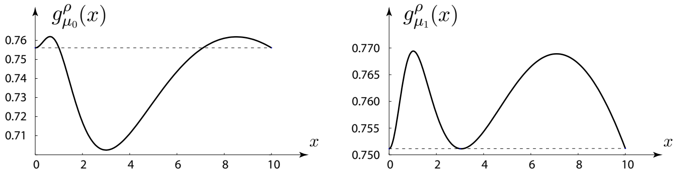

Fig. 10: The steepest decent cutting-plane algorithm algorithm converges in just two iterations in the normalized Rayleigh channel with $\rho = 0.5$ and no average power constraint.

The initial steps of the steepest ascent cutting plane algorithm require *very little computation* since the number of constraints and the number of variables is small. At the nth iteration, the alphabet contains at most $O(n)$ symbols, so that the complexity of the corresponding LP is polynomial in n [45].

The algorithm is convergent for models with finite peak power constraint: the proof is identical to the corresponding analysis of the steepest ascent cutting-plane algorithm for capacity calculation introduced in [37].

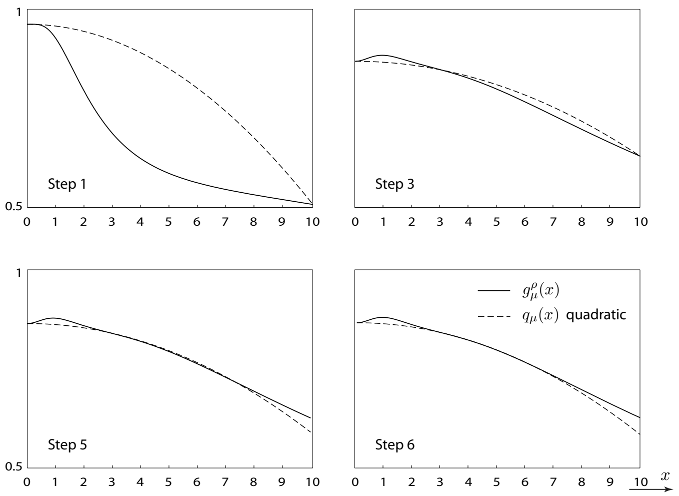

Fig. 11: Computation of the optimal input alphabet for the Rayleigh channel subject to both average and peak power constraints with $\rho = 0.5$.

The results shown in Figure 10 for the normalized Rayleigh channel were obtained using the steepest decent cutting-plane with $M = 10$ and $\sigma_P^2 = \infty$. The algorithm converged in just two iterations in this example. Results for the Rayleigh channel subject to both average and peak power constraints are shown in Figure 11 with $\sigma_P^2 = 10$ and $\rho = 0.5$. In this experiment the algorithm required about six iterations to compute μ^*.

IV. Conclusions

Many problems in information theory may be cast as a convex program over a set of probability distributions. Here we have seen three: hypothesis testing, channel capacity, and computation of the random coding exponent. Another example considered in [36] is computation of the distortion function in source coding. We have seen that the optimizing distribution is typically discrete [37], and in the case of the error exponent, this paper shows for the first time that the optimizer is *always discrete* when Assumptions (A1)–(A3) are satisfied.

Although the optimization problem in each case is infinite dimensional when the state space is not finite, in each example we have considered it is possible to construct a finite dimensional algorithm, and convergence is typically very fast. We believe this is in part due to the extremal nature of optimizers. Since optimizers have few points of support, this means the optimizer is on the boundary of the constraint set, and hence sensitivity is typically non-zero.

There are many interesting open questions:

(i) In this paper we restrict attention to the random coding exponent $E_r(R)$. However, for rates below the critical rate $R < R_{\text{crit}}$ there may be better bounds such as the sphere-packing bound, the expurgated bound, or the straight-line bound (see Section 5.8 of [30] and [52].) These bounds have a representation similar to the random coding exponent. Hence it is likely that analogous theory can be developed in these cases.

(ii) The theory described here sets the stage for further research on channel sensitivity. For example, how sensitive is the error exponent to SNR, coherence, channel memory, or other parameters?

(iii) It is of interest to see if efficient constellations can be adapted in a dynamic environment based on the cutting-plane algorithm.

(iv) In some applications optimal distributions may not be feasible due to the resulting ‘peakiness’ of the code book. There are then tradeoffs, and resource investment issues to be addressed. For example, one can consider if increasing the number of transmitter antennas will reduce the peakiness of the optimal random codebook.

(v) Finally, we are currently exploring extensions of the results reported here to MIMO models.

References

[1] I.C. Abou-Faycal, M.D. Trott, and S. Shamai. The capacity of discrete-time memoryless Rayleigh-fading channels. *IEEE Trans. Inform. Theory*, 47(4):1290–1301, 2001.

[2] V. Anantharam. A large deviations approach to error exponents in source coding and hypothesis testing. *IEEE Trans. Inform. Theory*, 36(4):938–943, 1990.

[3] S. Arimoto. Computation of random coding exponent functions. *IEEE Trans. Inform. Theory*, 22(6):665–671, 1976.

[4] E. A. Arutiunian. Bounds to the error probability exponent for semicontinuous memoryless channels. *Probl. Peredach. Inform.*, 4:37–48, 1968. (In Russian).

[5] R.R. Bahadur. *Some Limit Theorems in Statistics*. SIAM, Philadelphia, PA, 1971.

[6] Alexander Barg and G. David Forney, Jr. Random codes: minimum distances and error exponents. *IEEE Trans. Inform. Theory*, 48(9):2568–2573, 2002.

[7] Alexander Barg and Gilles Zémor. Error exponents of expander codes. *IEEE Trans. Inform. Theory*, 48(6):1725–1729, 2002. Special issue on Shannon theory: perspective, trends, and applications.

[8] Alexander Barg and Gilles Zémor. Error exponents of expander codes under linear-complexity decoding. *SIAM J. Discrete Math.*, 17(3):426–445 (electronic), 2004.

[9] A. Ben-Tal, M. Teboulle, and A. Charnes. The role of duality in optimization problems involving entropy functionals with applications to information theory. *J. Optim. Theory Appl.*, 58(2):209–223, 1988.

[10] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[11] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inform. Theory*, IT-20:405–417, 1974.

[12] R.E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, 18(4):460–473, 1972.

[13] R.E Blahut. *Principles and Practice of Information Theory*. McGraw-Hill, New York, 1995.

[14] J.M. Borwein and A.S. Lewis. A survey of convergence results for maximum entropy. In A. Mohammad-Djafari and G. Demoment, editors, *Maximum Entropy and Bayesian Methods*, pages 39–48. Kluwer Academic, Dordrecht, 1993.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[16] Terence H. Chan, Steve Hranilovic, and Frank R. Kschischang. Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs. *IEEE Trans. Inform. Theory*, 51(6):2073–2088, 2005.

[17] Rong-Rong Chen, B. Hajek, R. Koetter, and U. Madhow. On fixed input distributions for noncoherent communication over high SNR Rayleigh fading channels. *IEEE Trans. Inform. Theory*, 50(12):3390–3396, 2004.

[18] M. Chiang. Geometric programming for communication systems. *Foundations and Trends in Information and Communications Theory*, 2(1):1–156, 2005.

[19] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[20] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.

[21] I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, 1998. Information theory: 1948–1998.

[22] I. Csiszár and G. Longo. On the error exponent for source coding and for testing simple statistical hypotheses. *Studia Sci. Math. Hungar.*, 6:181–191, 1971.

[23] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.

[24] S.-C. Fang, J. R. Rajasekera, and H.-S. J. Tsao. *Entropy optimization and mathematical programming*. International Series in Operations Research & Management Science, 8. Kluwer Academic Publishers, Boston, MA, 1997.

[25] G. David Forney, Jr. *Concatenated codes*. The M.I.T. Press, Cambridge, Mass., 1966. M.I.T. Research Monograph, No. 37.

[26] G. David Forney, Jr. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Trans. Inform. Theory*, IT-14:206–220, 1968.

[27] R. G. Gallager. Low-density parity-check codes. *IRE Trans.*, IT-8:21–28, 1962.

[28] R. G. Gallager. *Low Density Parity Check Codes, Sc.D.* PhD thesis, Mass. Inst. Tech., Cambridge Mass, September, 1960.

[29] R.G. Gallager. Power limited channels: Coding, multiaccess, and spread spectrum. In R.E. Blahut and R. Koetter, editors, *Codes, Graphs, and Systems*, pages 229–257. Kluwer Academic Publishers, Boston, 2002.

[30] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

[31] Robert G. Gallager. The random coding bound is tight for the average code. *IEEE Trans. Inform. Theory*, IT-19(2):244–246, 1973.

[32] J.D. Gibson, R.L. Baker, T. Berger, T. Lookabaugh, and D. Lindbergh. *Digital Compression for Multimedia*. Morgan Kaufmann Publishers, San Fransisco, CA, 1998.

[33] M. Gursoy, H. V. Poor, and S. Verdú. The noncoherent Rician fading channel – Part I: Structure of the capacity achieving input. *IEEE Trans. Wireless Comm.*, 4(5):2193– 2206, 2005.

[34] M. Gursoy, H. V. Poor, and S. Verdú. The noncoherent Rician fading channel – Part II: Spectral efficiency in the low-power regime. *IEEE Trans. Wireless Comm.*, 4(5):2207– 2221, 2005.

[35] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, 36:369–408, 1965.

[36] J. Huang. *Characterization and computation of optimal distribution for channel coding*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2004.

[37] J. Huang and S. P. Meyn. Characterization and computation of optimal distribution for channel coding. *IEEE Trans. Inform. Theory*, 51(7):1–16, 2005.

[38] F. Jelinek. *Probabilistic Information Theory*. McGraw-Hill, NY, 1968.

[39] Hui Jin and Robert J. McEliece. Coding theorems for turbo code ensembles. *IEEE Trans. Inform. Theory*, 48(6):1451–1461, 2002. Special issue on Shannon theory: perspective, trends, and applications.

[40] M. Katz and S. Shamai. On the capacity-achieving distribution of the discrete-time non-coherent additive white gaussian noise channel. In *Proc. IEEE Int'l. Symp. Inform. Theory, Lausanne, Switzerland, June 30 - July 5.*, page 165, 2002.

[41] S. Kullback. *Information Theory and Statistics*. Dover Publications Inc., Mineola, NY, 1997. Reprint of the second (1968) edition.

[42] A. Lapidoth and N. Miliou. Duality bounds on the cut-off rate with applications to Ricean fading. to appear in IEEE Transactions on Information Theory, 2006.

[43] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. available from http://www.inference.phy.cam.ac.uk/mackay/itila/.

[44] G. Matz and P. Duhamel. Information geometric formulation and interpretation of accelerated blahut-arimoto-type algorithms. In *In Proc. IEEE Information Theory Workshop, San Antonio, TX*, pages 66–70, October 24-29 2004.

[45] N. Megiddo. On the complexity of linear programming. In T. F. Bewley, editor, *Advances in Economic Theory—The Fifth World Congress*, volume 12 of *Econometric Society Monographs*, pages 225–258. Cambridge University Press, Cambridge, UK, 1987.

[46] R. Palanki. On the capacity-achieving distributions of some fading channels. Presented at 40th Allerton Conference on Communication, Control, and Computing, 2002.

[47] C. Pandit. *Robust Statistical Modeling Based On Moment Classes With Applications to Admission Control, Large Deviations and Hypothesis Testing*. PhD thesis, University of Illinois at Urbana Champaign, University of Illinois, Urbana, IL, USA, 2004.

[48] C. Pandit and S. P. Meyn. Worst-case large-deviations with application to queueing and information theory. *Stoch. Proc. Applns.*, 116(5):724–756, 2006.

[49] C. Pandit, S. P. Meyn, and V. V. Veeravalli. Asymptotic robust Neyman-Pearson testing based on moment classes. In *Proceedings of the International Symposium on Information Theory (ISIT), 2004*, June 2004.

[50] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[51] S. Shamai and I. Bar-David. The capacity of average and peak-power-limited quadrature Gaussian channels. *IEEE Trans. Inform. Theory*, 41(4):1060–1071, 1995.

[52] Shlomo Shamai and Igal Sason. Variations on the Gallager bounds, connections, and applications. *IEEE Trans. Inform. Theory*, 48(12):3029–3051, 2002.

[53] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. II. *Information and Control*, 10:522–552, 1967.

[54] J. G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Inform. Contr.*, 18:203–219, 1971.

[55] IEEE Trans. Inform. Theory. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Trans. Inform. Theory*, 49(10):2426–2467, 2003.

[56] S. Verdu. On channel capacity per unit cost. *IEEE Trans. Inform. Theory*, 36(5):1019–1030, 1990.

[57] Ofer Zeitouni and Michael Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inform. Theory*, 37(2):285–290, 1991.